~~TOP SECRET EIDER~~

Data Storage and Data Retrieval Symposium
16-17 April 1959
Room 1W128, NSA, Ft. Meade
Sponsors: R/D, NSASAB

Dr. Kullback outlined the objectives of the symposium as exploring such questions as:

a. What is meant by a data storage and retrieval problem?
b. What are the present and prospective data storage and retrieval problems of NSA and its associated agencies?
c. What is the relative importance and interrelationship of these problems?

It was not to be an objective of this symposium to solve the problems raised, nor to explore data processing problems if storage and retrieval was not a major factor in them.

1. (Mr. Cram) In COSA (Collection and Signal Analysis) the major problems are:

a. Intercept Control. 12,000 cases, 6,000 assigned.
15,000 IBM cards, changes of 1,000 cards per week.
Monthly print of whole file.

Technical Extracts of Traffic Analysis

b. TEXTA: 500,000 IBM card file maintained, these cards are only partially fielded, but each card contains a particular type of information. These form an intermediate changeable storage from which punched tape can be made for outgoing messages, or complete prints for backing up new intercept stations.

c. Coverage Analysis Reports.

Hand Morse and Voice:	45,000 cards per month
Auto Morse.	150,000 cards per month
Radio Printer:	250,000 cards per month
Morse General Search:	250,000 cards per month
In prospect, Manual Morse:	5,000,000 cards per month

All cards are fielded, are transferred to M. T. to reduce space problems.

(b)(1)
(b)(3)-18 USC 798
(b)(3)-50 USC 3024(i)
(b)(3)-P.L. 86-36

d. [redacted]

e. Morse General Search: Several thousand items per week are sorted by frequency and distributed among 3 teams of people. Traffic which is identified as to country (about 75%) is sent to an analytic office. Unidentified traffic is punched on IBM cards, and stays at a level of about 23,000 items.

batches go out to [illegible]

f. Using IBM 650 for identifying all non-Russian callsigns, except Asiatic Communist.

Questions asked after the COSA presentation brought out the fact that AG handles most incoming traffic, amounting to 1.5 tons per day. While TEXTA file may contain information such as "What are all the transmitters in a certain region," it cannot practically answer the question because information is not all fielded. The point was made that we ought to record not enough data to answer all questions, but simply enough to answer the most important, most likely questions.

2. (Capt. Knepper) GENS has problems of data storage and retrieval in 3 areas.

a. Extraction of plain text materials from many files, contained in many forms. Access to millions of items is required for many individuals, some at remote locations from a central file. (This problem exists for the other 3 analytic offices as well).

b. Technical area

(1) Callsign identification (IBM-650 is in use on RU callsigns). System works, but needs larger storage and more than the 9 stations in 3 locations now available. (ADVA-2 also desired better access to central C/S identification).

(2) BIRDHOUSE Systems. numerous codebooks, strips, and strip positions. Central storage with inquiry stations is desired.

(3) SOAPFLAKES. Using IBM 704 to select wanted messages from a mass of plain text messages by means of word recognition. Depending upon the type of traffic, 5 to 90 percent of the messages is retained.

c. Management area: Keep records of intercept station assignments, performance, take, accessible to many potential users for direct inquiry.

Multiplicity of media compounds the GENS retrieval problem, one medium for which better processing means is wanted is plain speech, now recorded on audio tapes and manually transcribed for processing. Overall, want to get a single permanent record which can be read by machine and used for storage, and to get it as early as possible. (Other analytic offices stated the same need).

3. (Mr. Vergine, ADVA)

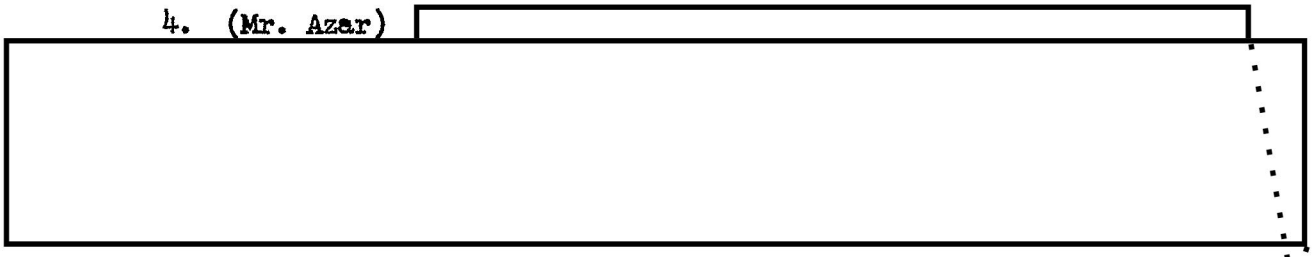
a. Magnetic Tape is being generated on SCRAMBLER at a rate of 14,000 tapes (45,000 intercept hours) monthly, and within a year will rise to 16,000 tapes (55,000 hours). Page print is also made which is the medium scanned for bust conditions. Automatic diarization (on-line sensing of particular events) is also being done, and may partially replace page scanning. Retrieval of this information stored on magnetic tape may take place as much as 3 to 5 years later. On some systems, up to 30% of the tapes can be disposed of, on other systems, we are keeping all.

b. CARP (Coverage Accounting Report, Printer)
Being used as a source of information for T/A in addition to its use by COSA for intercept accounting.

c. Mechanization of processing intelligence information. After getting information into machinable form and scanning the content as in SOAPFLAKES, would like to extract words which are of importance. Under the general heading of PERSON, PLACES, THINGS and TIME, would like to extract words and then relationships to each other. It is thought, for example, that 95% of the personality names which occur in text could be recognized by machine. Some few connectors (27) would display the general relationships which names, places and things have to each other.
Mechanical extraction from texts would support a library. The library would be capable of generating complete lists (people, commodities, etc.). It should be able to say generally what a particular day's transmission is "about." It should yield up original material on request. It should make associations of data, and analysts would be particularly interested in associations of which they were previously unaware.

In addition to emphasizing some of the points made by Capt.Knepper and Mr. Vergine, Mrs. Moody pointed out the need for a common heading format which could receive information as it develops and store it for future use. Also emphasized the need for early machinable recording of intercept, with elimination of repeated manual conversions.

4. (Mr. Azar)



(b)(1)
(b)(3)-18 USC 798
(b)(3)-50 USC 3024(i)
(b)(3)-P.L. 86-36

[redacted]

5. (Mr. Chesnut) [redacted]

6. (Mr. Byles) The data storage and retrieval problem of MPRO is largely that of storing data (in machine readable form) for other people which has once been machine processed, and is likely to be processed again. Precise identification to facilitate accurate retrieval is difficult. The following figures were presented.

a. Amount of material stored on various media.

FY	Cards	727 Tapes
57	78 million	6,052
58	100 million	9,300
59	72 million*	12,500*

FY	Cards	727 Tapes
60	55 million*	18,500*
61	50 million*	26,000*(equivalent to 320 TRACTOR tapes)
..		
65	30-40 million*	

* = estimate

About half of the 727 tapes are not used all the way down to the end, that is, they are not "stacked." TRACTOR tapes will have to be stacked and the contents of each reel will have to be internally identified on the tape. In order to have a unified file, 727 and card files may have to be similarly identified — a problem we have never solved in the past.

b. Age and use of files. Over a 2 month period, there were 77 calls for cards.

Age	% of file	% of calls
Less than 6 years old	25	55
Between 6 and 10 years old	50	20
Between 10 and 12 years old	20	21
More than 12 years old	5	4

Second Day of the Symposium

7. (Dr. Sanford) The Office of Central Reference provides information in support of COMINT, technical information to research, collects and organizes material to do both.

One million documents per year are selected from several million by CREF liaison representatives at reading panels throughout the intelligence community. In addition, each year 300,000 magazines, periodicals, journals, etc. are brought in. The yearly bill for publications is $180,000.

In support of COMINT we have accumulated 36 million selected, individual pages from documents. Each year we add another 2.6 million pages. From these files we answer 12,000 questions of the sort any library might answer and loan 36,000 books and 45,000 complete documents. In addition CREF gives spot answers to 120,000 questions posed by analysts who ask in the course of preparing a message for publication such questions as: "who was where, and when; what was the text of the agreement, what are the products of the factory at _____." The response time must

be short, because the requestor cannot wait for lengthy library research. In order to meet such questions, several specialized files exist:

a. PERSONS: 3 million people, half Russians, half from the rest of the world. Besides names, this is the place we record residence, occupation, affiliations, etc. and keep in effect a dossier of all we can get on these people.
b. THINGS. 100,000 commodity subject items.
c. PLACES. 500,000 place names, and associated information such as organizations and people in charge, manufactures and stockpiles, transportation, tec.
d. PROCESSES. 100,000 items.

Although some experiments exist on mechanizing reference material, virtually all our work is done manually, with the result that document dissemination requires only 25% of the resources of CREF, and the other 75% is consumed in the cataloging, filing and finding operations. Mechanization of our files is a real challenge. The 36 million items we now file as pages if put on magnetic tape might involve the storage of some 180 billion bits.

There are three classes of ways in which we could handle our information.

a. Keep only one copy, and use a classification system with cross references. Although there are situations in which this approach works, it does not always work for us.
b. Make multiple files, anticipating future use. Each page may go in several times: once into a PERSON file, several times into organization, commodity, or place file. This is the way we operate now.
c. Keep only one copy, but have a multiplicity of ways to get at it, "fishhooks" by which the contents can be caught. Many modern systems use this approach under such names as uniterms or descriptors.

This last way might give us an excellent reduction in material required to be stored our yearly increment of 2.6 million pages might be as low as .4 million under this system. Some 17 "fishhooks" might be the average number required per page. We have already found that on a 100,000 document collection, we can retrieve any document with an average of 9.5 fishhooks. Although we have proved that this coordinate indexing scheme works, it is not mechanized, and it is very slow.

As with the cryppies and their intercept material, reference services need to get their material in some machinable form, and the earlier it is done, the better for all. There is some danger to NSA that someone else will start the ball rolling on recording or processing in a direction in which we don't want it to go. We have a splendid chance to mechanize in the reference area, many gains to be made, and possibly a great deal to lose if we don't.

8. Discussion, Friday.

(Dr. Ridenour) Too little attention has been given to the fact that someone will have to make the initial machine recordings. Modern systems offer excellent media for reducing the bulk of files. They also offer extremely high scanning speeds. Abstracting and cataloguing are expensive, since they require well-trained people whose pace cannot be rapid. Perhaps it is now possible to avoid fine cataloguing and abstracting if a file can be read exhaustively in a reasonable time.

(Dr. Weaver) Machine scanning speeds of 1 to 100 million per second, when contrasted with human speeds of the order of 1 second, suggest that the difference is not merely quantitative. A qualitative difference exists, which permits a new approach. Even such speeds, however, may not make much impact if the machine must approach problems in a very formal way, without the associative power of the human mind. Some basic research needs to be done to understand how people get trained in logic, linguistics, etc. We have to know the relationships of words and ideas.

(Dr. Tukey) Perhaps we should focus on phrases — the larger entities in which ideas are expressed — not on words.

(Dr. Jacobs) For all of the material which we handle — be it traffic, information about traffic, or collateral information — we need descriptive statements in three areas: volume, speed, and manner of selection. We have been reminded that volume will increase rather than diminish, especially as we try to process material that we do very little with now. Speed, perhaps insufficiently emphasized during this symposium, will need to be increased. We have discussed the selection problem only indirectly. We will always be uncovering new material, and will not be able to save everything. No matter what our system, we will not be able to wade through all the material we can collect, generate, or think of generating. By selection I mean finding out what is essential because you must get rid of what is not essential. There was much similarity in the problems described by the four analytic offices. We can no longer tolerate special solutions to similar problems. That problems are common or even similar encourages hope that integrated solutions can alleviate these problems at the same time with the same approach.

(Mr. McPherson) There seem to be two broad kinds of retrieval—direct and associative

a. Direct retrieval

(1) Ask for information by the pocket in which it is stored, rather than by the information itself. The simplest case of this kind of retrieval is where all of the data is withdrawn from storage in the order in which it was put in.
(2) Records (or requests) of doubtful completeness or accuracy. Finding a man's name (in a name file) when the spelling is doubtful.
(3) Arbitrary selective retrieval — finding a needle in a haystack. Pick one man's activities on a particular day from hundreds of thousands of pieces of data. While this may still be direct, this last problem is complicated by lack of an efficient means of getting through a large volume of simply identified information.

b. Associative retrieval, in which information is found only by secondary information, not in the order in which it was originally arranged, nor by the items by which it was originally identified. A large volume of information will have to be scanned in some search fashion.

There are other problems, in addition to retrieval per se, which ought to be considered at the same time. One is that of condensing information, if we could boil information down into Persons, Places, Things and Time before we declare it ready for retrieval purposes, it might be more accessible. The other problem is that of in-filing and out-filing. If you have to wait to insert new information in its proper place, and use this opportunity to delete old information, there may be a conflict with the need for timely information.

9. (Mr. Page and Mrs. Peifer) Data Problems in CSEC

a. Reference library of COMSEC documents 300,000 pages, average document size 30 pp., with a range of 1 to 150 pp. Access of 10 minutes to 2 days is required, the detailed portion of the document wanted is not known.
b. Reference library of key cards. Card required can be specified, must be found in from 10 minutes to a half-day in a bank which will range from about 100,000 cards next year to half a million cards in 5 years.

c. Records for control of COMSEC materials. The security of COMSEC material depends on an effective system of custody and accountability. COMSEC material is controlled by short title and serial number through 5 types of actions production, initial accountability, initial issue, lateral transfer, and destruction.

The Control Division is involved in at least two of these five types of actions — initial accountability and initial issue. There are 100 million items now in the system, and another 12 million are added each year. Each item may be involved in a number of transactions, lateral transfer of hardware items, for example, might occur a number of times. Present records, however, are not as numerous as the items accounted for. For example, one record may be made of the receipt of 5,000 code books from Production. As these are issued, a half-dozen records might be made.

The activity of a year in the record file might be approximately as follows

Annual record entries (queries)	Reason	Retrieval Time
25,000	12 million newly produced items	current
150,000	initial issue of new material	current
134,000	lateral transfers	current
183,000	destruction	current
492,000	reflect changes in item file	
800,000	inventory	
1,292,000	entries to account for items	
1,000	queries, possible compromise	2 hours
500	queries, manual changes, etc.	2 hours
1,300	queries, miscellaneous distribution	1-3 days
2,800	unscheduled queries, with quick response required.	

Records have been in written form, manually maintained. They are now being converted to IBM cards for processing on standard EAM. The number of columns per record is about 60.

In addition to the record file, there is associated information about COMSEC material which will also be put into an IBM card system. This will allow machine search for information, and machine preparation of forms accompanying material. About 10,000 basic short titles will be included, and will occupy about 100,000 cards.